PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : C12N 15/53, 15/82, 9/02, C12Q 1/68, 1/02 | A2

(11) International Publication Number: WO 00/37652

(43) International Publication Date: 29 June 2000 (29.06.00)

(21) International Application Number: PCT/US99/30337

(22) International Filing Date: 20 December 1999 (20.12.99)

(30) Priority Data:
60/113,190 21 December 1998 (21.12.98) US

(71) Applicant *(for all designated States except US)*: E.I. DU PONT DE NEMOURS AND COMPANY [US/US]; 1007 Market Street, Wilmington, DE 19898 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: FAMODU, Omolayo, O. [US/US]; 216 Barrett Run Place, Newark, DE 19702 (US). MCGONIGLE, Brian [US/US]; 1707 North Union Street, Wilmington, DE 19806 (US). ODELL, Joan, T. [US/US]; P.O. Box 826, Unionville, PA 19375 (US). FADER, Gary, M. [US/US]; 1000 Woods Lane, Landenberg, PA 19350 (US). FALCO, Saverio, Carl [US/US]; 1902 Miller Road, Arden, DE 19810 (US).

(74) Agent: FEULNER, Gregory, J.; E.I. du Pont de Nemours and Company, Legal Patent Record Center, 1007 Market Street, Wilmington, DE 19898 (US).

(81) Designated States: AE, AL, AU, BA, BB, BG, BR, CA, CN, CR, CU, CZ, DM, EE, GD, GE, HR, HU, ID, IL, IN, IS, JP, KP, KR, LC, LK, LR, LT, LV, MG, MK, MN, MX, NO, NZ, PL, RO, SG, SI, SK, SL, TR, TT, UA, US, UZ, VN, YU, ZA, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: FLAVONOID BIOSYNTHETIC ENZYMES

(57) Abstract

This invention relates to an isolated nucleic acid fragment encoding a flavonoid biosynthetic enzyme. The invention also relates to the construction of a chimeric gene encoding all or a portion of the flavonoid biosynthetic enzyme, in sense or antisense orientation, wherein expression of the chimeric gene results in production of altered levels of the flavonoid biosynthetic enzyme in a transformed host cell.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

TITLE

FLAVONOID BIOSYNTHETIC ENZYMES

This application claims the benefit of U.S. Provisional Application No. 60/113,190, filed December 21, 1998.

FIELD OF THE INVENTION

This invention is in the field of plant molecular biology. More specifically, this invention pertains to nucleic acid fragments encoding flavonoid biosynthetic enzymes in plants and seeds.

BACKGROUND OF THE INVENTION

Plants accumulate a variety of natural products that are synthesized in response to environmental stimuli and genetically programmed developmental signals. Of these secondary metabolites, flavonoids are probably the most ubiquitous. Flavonoids are products of biosynthetic pathways originating from the central phenylpropanoid pathway and the acetate-malonate pathway. Flavonoids represent a large class of molecules that serve many diverse functions, such as co-pigments in flower color, stimulation of pollen tube growth, pollinator attraction, and feeding deterrents and protection against UV irradiation in fruits and seeds (Holton, T. A. et al. (1993) *Plant J 4*:1003-1010). Therefore, there has been considerable interest in understanding flavonoid biosynthesis, which in turn would afford the ability to control pigmentation and many other phenotypic traits associated with flavonoid synthesis.

Isoflavaonoids constitute a large group of naturally occurring flavonoids. They occur almost exclusively in leguminous plants and form part of the host defense response against phytopathogenic microorganisms (Akashi, T. et al. (1998) *Biochemical and Biophysical Rseaerch Comunications 251*:67-70). For example, a variety of simple isoflavonoids like coumestans, pterocarpans and isoflavans have been shown to have antimicrobial properties.

An essential part of the biosynthesis of these compounds is the 2'-hydroxylation of isoflavones by isoflavone 2'-hydroxylase (I2'H). It may be possible to change the levels of some isoflavones in plants by altering the level or activity of isoflavone biosynthetic enzymes. For example, one unique isoflavone, daidzein (5-deoxyisoflavonoid), is the product of the phenylpropanoid pathway and is a healthful compound found in soybean seed. Isoflavone O-methyltransference, catalyzed by isoflavone-O-methyltransferase, is the first step in degradation of daidzein. It may be possible to suppress the activity of this enzyme causing daidzein concentrations to increase yielding higher levels of this beneficial isoflavone in soybean seed.

Some flavonoids have sulfate groups attached to common flavones and flavonols via ester linkages. For example, flavonol 3-sulfotransferase and flavonol 4-sulfotransferase catalyze the stepwise sulfation of quercetin to form quercetin 3,4'-disulfate. The functional significance of flavonoid sulfates in plant tissue is not clear, however, they may play a role

in the detoxification of reactive hydroxyl groups, and/or may be an important molecule for sequestering sulfate ions (Akashi, T. et al. (1998) *Biochemical and Biophysical Research Communications 251*:67-70).

There is a great deal of interest in identifying the genes that encode proteins involved in flavonoid biosynthesis in plants. The genes that code for these enzymes may be used to study the interactions among individuals of the pathways and develop methods to modulate the anthocyanin and flavonol biosynthetic pathways to control flavonoid biosynthesis. Accordingly, the availability of nucleic acid sequences encoding all or a portion of a flavonoid biosynthetic enzyme would facilitate studies to better understand flavonol biosynthesis in plants and provide genetic tools to enhance or otherwise alter flavonol and anthocyanin biosynthesis.

SUMMARY OF THE INVENTION

The present invention relates to isolated polynucleotides comprising a nucleotide sequence encoding a polypeptide of at least 494 amino acids that has at least 80% identity based on the Clustal method of alignment when compared to a polypeptide selected from the group consisting of soybean isoflavone 2-hydroxylase polypeptides of SEQ ID NO:2 and 4. The present invention also relates to an isolated polynucleotide comprising the complement of the nucleotide sequences described above.

The isolated polynucleotides of the claimed invention may consist of a nucleic acid sequence selected from the group consisting of SEQ ID NOs:1, 3 and 5 that codes for the polypeptide selected from the group consisting of SEQ ID NOs:2, 4 and 6. The present invention also relates to an isolated polynucleotide comprising a nucleotide sequences of at least one of 60 (preferably at least one of 40, most preferably at least one of 30) contiguous nucleotides derived from a nucleotide sequence selected from the group consisting of SEQ ID NOs:1, 3, 5 and the complement of such nucleotide sequences.

The present invention relates to a chimeric gene comprising an isolated polynucleotide of the present invention operably linked to suitable regulatory sequences.

The present invention relates to an isolated host cell comprising a chimeric gene of the present invention or an isolated polynucleotide of the present invention. The host cell may be eukaryotic, such as a yeast or a plant cell, or prokaryotic, such as a bacterial cell. The present invention also relates to a virus, preferably a baculovirus, comprising an isolated polynucleotide of the present invention or a chimeric gene of the present invention.

The present invention relates to a process for producing an isolated host cell comprising a chimeric gene of the present invention or an isolated polynucleotide of the present invention, the process comprising either transforming or transfecting an isolated compatible host cell with a chimeric gene or isolated polynucleotide of the present invention.

The present invention relates to a isoflavone 2-hydroxylase polypeptide of at least 494 amino acids comprising at least 80% homology based on the Clustal method of alignment compared to a polypeptide selected from the group consisting of SEQ ID NOs:2 and 4.

The present invention relates to a method of selecting an isolated polynucleotide that affects the level of expression of an isoflavone 2-hydroxylase polypeptide in a host cell, preferably a plant cell, the method comprising the steps of: (a) constructing an isolated polynucleotide of the present invention or an isolated chimeric gene of the present invention; (b) introducing the isolated polynucleotide or the isolated chimeric gene into a host cell; (c) measuring the level a isoflavone 2-hydroxylase polypeptide in the host cell containing the isolated polynucleotide; and (d) comparing the level of a isoflavone 2-hydroxylase polypeptide in the host cell containing the isolated polynucleotide with the level of an isoflavone 2-hydroxylase polypeptide in the host cell that does not contain the isolated polynucleotide.

The present invention relates to a method of obtaining a nucleic acid fragment encoding a substantial portion of an isoflavone 2-hydroxylase polypeptide gene, preferably a plant isoflavone 2-hydroxylase polypeptide gene, comprising the steps of: synthesizing an oligonucleotide primer comprising a nucleotide sequence of at least one of 60 (preferably at least one of 40, most preferably at least one of 30) contiguous nucleotides derived from a nucleotide sequence selected from the group consisting of SEQ ID NOs:1 and 3 and the complement of such nucleotide sequences; and amplifying a nucleic acid fragment (preferably a cDNA inserted in a cloning vector) using the oligonucleotide primer. The amplified nucleic acid fragment preferably will encode a portion of an isoflavone 2-hydroxylase amino acid sequence.

The present invention also relates to a method of obtaining a nucleic acid fragment encoding all or a substantial portion of the amino acid sequence encoding an isoflavone 2-hydroxylase polypeptide comprising the steps of: probing a cDNA or genomic library with an isolated polynucleotide of the present invention; identifying a DNA clone that hybridizes with an isolated polynucleotide of the present invention; isolating the identified DNA clone; and sequencing the cDNA or genomic fragment that comprises the isolated DNA clone.

The present invention relates to a composition, such as a hybridization mixture, comprising an isolated polynucleotide of the present invention.

The present invention relates to an isolated polynucleotide of the present invention comprising at least one of 30 contiguous nucleotides derived from a nucleic acid sequence selected from the group consisting of SEQ ID NOs:1 and 3.

The present invention relates to an expression cassette comprising an isolated polynucleotide of the present invention operably linked to a promoter.

The present invention relates to a method for positive selection of a transformed cell comprising: (a) transforming a host cell with the chimeric gene of the present invention or

an expression cassette of the present invention; and (b) growing the transformed host cell, preferably plant cell, such as a monocot or a dicot, under conditions which allow expression of the isoflavone 2-hydroxylase polynucleotide in an amount sufficient to complement a mutant cell with altered isoflavone 2-hydroxylase activity to provide a positive selection means.

BRIEF DESCRIPTION OF THE SEQUENCE DESCRIPTIONS

The invention can be more fully understood from the following detailed description and the accompanying Sequence Listing which form a part of this application.

Table 1 lists the polypeptides that are described herein, the designation of the cDNA clones that comprise the nucleic acid fragments encoding polypeptides representing all or a substantial portion of these polypeptides, and the corresponding identifier (SEQ ID NO:) as used in the attached Sequence Listing. Table 1 also identifies the cDNA clones as individual ESTs ("EST"), the sequences of the entire cDNA inserts comprising the indicated cDNA clones ("FIS"), contigs assembled from two or more ESTs ("Contig"), contigs assembled from an FIS and one or more ESTs ("Contig*"), or sequences encoding the entire protein derived from an FIS, a contig, or an FIS and PCR ("CGS"). Nucleotide sequences, SEQ ID NOs:1 and 3and amino acid sequences SEQ ID NOs:2 and 4 were determined by further sequence analysis of cDNA clone encoding the amino acid sequence set forth in SEQ ID NO:6. Nucleotide SEQ ID NO:5 and amino acid SEQ ID NO:6 were presented in a U.S. Provisional Application No. 60/113,190, filed December 21, 1998.

The sequence descriptions and Sequence Listing attached hereto comply with the rules governing nucleotide and/or amino acid sequence disclosures in patent applications as set forth in 37 C.F.R. §1.821-1.825.

TABLE 1

Flavonoid Biosynthetic Enzymes

Protein	Clone Designation	SEQ ID NO: (Nucleotide)	SEQ ID NO: (Amino Acid)
Isoflavone 2-hydroxylase	sls1c.pk005.n3 (FIS)	1	2
Isoflavone 2-hydroxylase	src3c.pk005.f5 (FIS)	3	4
Isoflavone 2-hydroxylase	Contig composed of: sgc1c.pk001.g17 sgs2c.pk004.h7 sfl1.pk0034.g1	5	6

The Sequence Listing contains the one letter code for nucleotide sequence characters and the three letter codes for amino acids as defined in conformity with the IUPAC-IUBMB standards described in *Nucleic Acids Res. 13*:3021-3030 (1985) and in the *Biochemical J. 219 (No. 2)*:345-373 (1984) which are herein incorporated by reference. The symbols and

format used for nucleotide and amino acid sequence data comply with the rules set forth in 37 C.F.R. §1.822.

DETAILED DESCRIPTION OF THE INVENTION

In the context of this disclosure, a number of terms shall be utilized. As used herein, a "polynucleotide" is a nucleotide sequence such as a nucleic acid fragment. A polynucleotide may be a polymer of RNA or DNA that is single- or double-stranded, that optionally contains synthetic, non-natural or altered nucleotide bases. A polynucleotide in the form of a polymer of DNA may be comprised of one or more segments of cDNA, genomic DNA, synthetic DNA, or mixtures thereof. An isolated polynucleotide of the present invention may include at least one of 60 contiguous nucleotides, preferably at least one of 40 contiguous nucleotides, most preferably one of at least 30 contiguous nucleotides derived from SEQ ID NOs:1, 3, 5 or the complement of such sequences.

As used herein, "contig" refers to a nucleotide sequence that is assembled from two or more constituent nucleotide sequences that share common or overlapping regions of sequence homology. For example, the nucleotide sequences of two or more nucleic acid fragments can be compared and aligned in order to identify common or overlapping sequences. Where common or overlapping sequences exist between two or more nucleic acid fragments, the sequences (and thus their corresponding nucleic acid fragments) can be assembled into a single contiguous nucleotide sequence.

As used herein, "substantially similar" refers to nucleic acid fragments wherein changes in one or more nucleotide bases results in substitution of one or more amino acids, but do not affect the functional properties of the polypeptide encoded by the nucleotide sequence. "Substantially similar" also refers to nucleic acid fragments wherein changes in one or more nucleotide bases does not affect the ability of the nucleic acid fragment to mediate alteration of gene expression by gene silencing through for example antisense or co-suppression technology. "Substantially similar" also refers to modifications of the nucleic acid fragments of the instant invention such as deletion or insertion of one or more nucleotides that do not substantially affect the functional properties of the resulting transcript vis-à-vis the ability to mediate gene silencing or alteration of the functional properties of the resulting protein molecule. It is therefore understood that the invention encompasses more than the specific exemplary nucleotide or amino acid sequences and includes functional equivalents thereof.

Substantially similar nucleic acid fragments may be selected by screening nucleic acid fragments representing subfragments or modifications of the nucleic acid fragments of the instant invention, wherein one or more nucleotides are substituted, deleted and/or inserted, for their ability to affect the level of the polypeptide encoded by the unmodified nucleic acid fragment in a plant or plant cell. For example, a substantially similar nucleic acid fragment representing at least one of 30 contiguous nucleotides derived from the instant nucleic acid

fragment can be constructed and introduced into a plant or plant cell. The level of the polypeptide encoded by the unmodified nucleic acid fragment present in a plant or plant cell exposed to the substantially similar nucleic fragment can then be compared to the level of the polypeptide in a plant or plant cell that is not exposed to the substantially similar nucleic acid fragment.

For example, it is well known in the art that antisense suppression and co-suppression of gene expression may be accomplished using nucleic acid fragments representing less than the entire coding region of a gene, and by nucleic acid fragments that do not share 100% sequence identity with the gene to be suppressed. Moreover, alterations in a nucleic acid fragment which result in the production of a chemically equivalent amino acid at a given site, but do not effect the functional properties of the encoded polypeptide, are well known in the art. Thus, a codon for the amino acid alanine, a hydrophobic amino acid, may be substituted by a codon encoding another less hydrophobic residue, such as glycine, or a more hydrophobic residue, such as valine, leucine, or isoleucine. Similarly, changes which result in substitution of one negatively charged residue for another, such as aspartic acid for glutamic acid, or one positively charged residue for another, such as lysine for arginine, can also be expected to produce a functionally equivalent product. Nucleotide changes which result in alteration of the N-terminal and C-terminal portions of the polypeptide molecule would also not be expected to alter the activity of the polypeptide. Each of the proposed modifications is well within the routine skill in the art, as is determination of retention of biological activity of the encoded products. Consequently, an isolated polynucleotide comprising a nucleotide sequence of at least one of 60 (preferably at least one of 40, most preferably at least one of 30) contiguous nucleotides derived from a nucleotide sequence selected from the group consisting of SEQ ID NOs:1 and 3 and the complement of such nucleotide sequences may be used in methods of selecting an isolated polynucleotide that affects the expression of a polypeptide (isoflavone 2-hydroxylase) in a host cell. A method of selecting an isolated polynucleotide that affects the level of expression of a polypeptide in a host cell (eukaryotic, such as plant or yeast, prokaryotic such as bacterial, or viral) may comprise the steps of: constructing an isolated polynucleotide of the present invention or an isolated chimeric gene of the present invention; introducing the isolated polynucleotide or the isolated chimeric gene into a host cell; measuring the level a polypeptide in the host cell containing the isolated polynucleotide; and comparing the level of a polypeptide in the host cell containing the isolated polynucleotide with the level of a polypeptide in a host cell that does not contain the isolated polynucleotide.

Moreover, substantially similar nucleic acid fragments may also be characterized by their ability to hybridize. Estimates of such homology are provided by either DNA-DNA or DNA-RNA hybridization under conditions of stringency as is well understood by those skilled in the art (Hames and Higgins, Eds. (1985) Nucleic Acid Hybridisation, IRL Press,

Oxford, U.K.). Stringency conditions can be adjusted to screen for moderately similar fragments, such as homologous sequences from distantly related organisms, to highly similar fragments, such as genes that duplicate functional enzymes from closely related organisms. Post-hybridization washes determine stringency conditions. One set of preferred conditions uses a series of washes starting with 6X SSC, 0.5% SDS at room temperature for 15 min, then repeated with 2X SSC, 0.5% SDS at 45°C for 30 min, and then repeated twice with 0.2X SSC, 0.5% SDS at 50°C for 30 min. A more preferred set of stringent conditions uses higher temperatures in which the washes are identical to those above except for the temperature of the final two 30 min washes in 0.2X SSC, 0.5% SDS was increased to 60°C. Another preferred set of highly stringent conditions uses two final washes in 0.1X SSC, 0.1% SDS at 65°C.

Substantially similar nucleic acid fragments of the instant invention may also be characterized by the percent identity of the amino acid sequences that they encode to the amino acid sequences disclosed herein, as determined by algorithms commonly employed by those skilled in this art. Suitable nucleic acid fragments (isolated polynucleotides of the present invention) encode polypeptides that are at least about 70% identical, preferably at least about 80% identical to the amino acid sequences reported herein. Preferred nucleic acid fragments encode amino acid sequences that are about 85% identical to the amino acid sequences reported herein. More preferred nucleic acid fragments encode amino acid sequences that are at least about 90% identical to the amino acid sequences reported herein. Most preferred are nucleic acid fragments that encode amino acid sequences that are at least about 95% identical to the amino acid sequences reported herein. Suitable nucleic acid fragments not only have the above homologies but typically encode a polypeptide having at least about 50 amino acids, preferably at least about 100 amino acids, more preferably at least about 150 amino acids, still more preferably at least about 200 amino acids, and most preferably at least about 250 amino acids. Sequence alignments and percent identity calculations were performed using the Megalign program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). Multiple alignment of the sequences was performed using the Clustal method of alignment (Higgins and Sharp (1989) *CABIOS. 5*:151-153) with the default parameters (GAP PENALTY=10, GAP LENGTH PENALTY=10). Default parameters for pairwise alignments using the Clustal method were KTUPLE 1, GAP PENALTY=3, WINDOW=5 and DIAGONALS SAVED=5.

A "substantial portion" of an amino acid or nucleotide sequence comprises an amino acid or a nucleotide sequence that is sufficient to afford putative identification of the protein or gene that the amino acid or nucleotide sequence comprises. Amino acid and nucleotide sequences can be evaluated either manually by one skilled in the art, or by using computer-based sequence comparison and identification tools that employ algorithms such as BLAST (Basic Local Alignment Search Tool; Altschul et al. (1993) *J. Mol. Biol. 215*:403-410; see

also www.ncbi.nlm.nih.gov/BLAST/). In general, a sequence of ten or more contiguous amino acids or thirty or more contiguous nucleotides is necessary in order to putatively identify a polypeptide or nucleic acid sequence as homologous to a known protein or gene. Moreover, with respect to nucleotide sequences, gene-specific oligonucleotide probes comprising 30 or more contiguous nucleotides may be used in sequence-dependent methods of gene identification (e.g., Southern hybridization) and isolation (e.g., *in situ* hybridization of bacterial colonies or bacteriophage plaques). In addition, short oligonucleotides of 12 or more nucleotides may be used as amplification primers in PCR in order to obtain a particular nucleic acid fragment comprising the primers. Accordingly, a "substantial portion" of a nucleotide sequence comprises a nucleotide sequence that will afford specific identification and/or isolation of a nucleic acid fragment comprising the sequence. The instant specification teaches amino acid and nucleotide sequences encoding polypeptides that comprise one or more particular plant proteins. The skilled artisan, having the benefit of the sequences as reported herein, may now use all or a substantial portion of the disclosed sequences for purposes known to those skilled in this art. Accordingly, the instant invention comprises the complete sequences as reported in the accompanying Sequence Listing, as well as substantial portions of those sequences as defined above.

"Codon degeneracy" refers to divergence in the genetic code permitting variation of the nucleotide sequence without effecting the amino acid sequence of an encoded polypeptide. Accordingly, the instant invention relates to any nucleic acid fragment comprising a nucleotide sequence that encodes all or a substantial portion of the amino acid sequences set forth herein. The skilled artisan is well aware of the "codon-bias" exhibited by a specific host cell in usage of nucleotide codons to specify a given amino acid. Therefore, when synthesizing a nucleic acid fragment for improved expression in a host cell, it is desirable to design the nucleic acid fragment such that its frequency of codon usage approaches the frequency of preferred codon usage of the host cell.

"Synthetic nucleic acid fragments" can be assembled from oligonucleotide building blocks that are chemically synthesized using procedures known to those skilled in the art. These building blocks are ligated and annealed to form larger nucleic acid fragments which may then be enzymatically assembled to construct the entire desired nucleic acid fragment. "Chemically synthesized", as related to nucleic acid fragment, means that the component nucleotides were assembled *in vitro*. Manual chemical synthesis of nucleic acid fragments may be accomplished using well established procedures, or automated chemical synthesis can be performed using one of a number of commercially available machines. Accordingly, the nucleic acid fragments can be tailored for optimal gene expression based on optimization of nucleotide sequence to reflect the codon bias of the host cell. The skilled artisan appreciates the likelihood of successful gene expression if codon usage is biased towards

those codons favored by the host. Determination of preferred codons can be based on a survey of genes derived from the host cell where sequence information is available.

"Gene" refers to a nucleic acid fragment that expresses a specific protein, including regulatory sequences preceding (5' non-coding sequences) and following (3' non-coding sequences) the coding sequence. "Native gene" refers to a gene as found in nature with its own regulatory sequences. "Chimeric gene" refers any gene that is not a native gene, comprising regulatory and coding sequences that are not found together in nature. Accordingly, a chimeric gene may comprise regulatory sequences and coding sequences that are derived from different sources, or regulatory sequences and coding sequences derived from the same source, but arranged in a manner different than that found in nature. "Endogenous gene" refers to a native gene in its natural location in the genome of an organism. A "foreign" gene refers to a gene not normally found in the host organism, but that is introduced into the host organism by gene transfer. Foreign genes can comprise native genes inserted into a non-native organism, or chimeric genes. A "transgene" is a gene that has been introduced into the genome by a transformation procedure.

"Coding sequence" refers to a nucleotide sequence that codes for a specific amino acid sequence. "Regulatory sequences" refer to nucleotide sequences located upstream (5' non-coding sequences), within, or downstream (3' non-coding sequences) of a coding sequence, and which influence the transcription, RNA processing or stability, or translation of the associated coding sequence. Regulatory sequences may include promoters, translation leader sequences, introns, and polyadenylation recognition sequences.

"Promoter" refers to a nucleotide sequence capable of controlling the expression of a coding sequence or functional RNA. In general, a coding sequence is located 3' to a promoter sequence. The promoter sequence consists of proximal and more distal upstream elements, the latter elements often referred to as enhancers. Accordingly, an "enhancer" is a nucleotide sequence which can stimulate promoter activity and may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue-specificity of a promoter. Promoters may be derived in their entirety from a native gene, or be composed of different elements derived from different promoters found in nature, or even comprise synthetic nucleotide segments. It is understood by those skilled in the art that different promoters may direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental conditions. Promoters which cause a nucleic acid fragment to be expressed in most cell types at most times are commonly referred to as "constitutive promoters". New promoters of various types useful in plant cells are constantly being discovered; numerous examples may be found in the compilation by Okamuro and Goldberg (1989) *Biochemistry of Plants 15*:1-82. It is further recognized that since in most cases the exact boundaries of regulatory sequences

have not been completely defined, nucleic acid fragments of different lengths may have identical promoter activity.

The "translation leader sequence" refers to a nucleotide sequence located between the promoter sequence of a gene and the coding sequence. The translation leader sequence is present in the fully processed mRNA upstream of the translation start sequence. The translation leader sequence may affect processing of the primary transcript to mRNA, mRNA stability or translation efficiency. Examples of translation leader sequences have been described (Turner and Foster (1995) *Mol. Biotechnol. 3*:225-236).

The "3' non-coding sequences" refer to nucleotide sequences located downstream of a coding sequence and include polyadenylation recognition sequences and other sequences encoding regulatory signals capable of affecting mRNA processing or gene expression. The polyadenylation signal is usually characterized by affecting the addition of polyadenylic acid tracts to the 3' end of the mRNA precursor. The use of different 3' non-coding sequences is exemplified by Ingelbrecht et al. (1989) *Plant Cell 1*:671-680.

"RNA transcript" refers to the product resulting from RNA polymerase-catalyzed transcription of a DNA sequence. When the RNA transcript is a perfect complementary copy of the DNA sequence, it is referred to as the primary transcript or it may be a RNA sequence derived from posttranscriptional processing of the primary transcript and is referred to as the mature RNA. "Messenger RNA (mRNA)" refers to the RNA that is without introns and that can be translated into polypeptide by the cell. "cDNA" refers to a double-stranded DNA that is complementary to and derived from mRNA. "Sense" RNA refers to an RNA transcript that includes the mRNA and so can be translated into a polypeptide by the cell. "Antisense RNA" refers to an RNA transcript that is complementary to all or part of a target primary transcript or mRNA and that blocks the expression of a target gene (see U.S. Patent No. 5,107,065, incorporated herein by reference). The complementarity of an antisense RNA may be with any part of the specific nucleotide sequence, i.e., at the 5' non-coding sequence, 3' non-coding sequence, introns, or the coding sequence. "Functional RNA" refers to sense RNA, antisense RNA, ribozyme RNA, or other RNA that may not be translated but yet has an effect on cellular processes.

The term "operably linked" refers to the association of two or more nucleic acid fragments on a single nucleic acid fragment so that the function of one is affected by the other. For example, a promoter is operably linked with a coding sequence when it is capable of affecting the expression of that coding sequence (i.e., that the coding sequence is under the transcriptional control of the promoter). Coding sequences can be operably linked to regulatory sequences in sense or antisense orientation.

The term "expression", as used herein, refers to the transcription and stable accumulation of sense (mRNA) or antisense RNA derived from the nucleic acid fragment of the invention. Expression may also refer to translation of mRNA into a polypeptide.

"Antisense inhibition" refers to the production of antisense RNA transcripts capable of suppressing the expression of the target protein. "Overexpression" refers to the production of a gene product in transgenic organisms that exceeds levels of production in normal or non-transformed organisms. "Co-suppression" refers to the production of sense RNA transcripts capable of suppressing the expression of identical or substantially similar foreign or endogenous genes (U.S. Patent No. 5,231,020, incorporated herein by reference).

"Altered levels" refers to the production of gene product(s) in transgenic organisms in amounts or proportions that differ from that of normal or non-transformed organisms.

"Mature" protein refers to a post-translationally processed polypeptide; i.e., one from which any pre- or propeptides present in the primary translation product have been removed. "Precursor" protein refers to the primary product of translation of mRNA; i.e., with pre- and propeptides still present. Pre- and propeptides may be but are not limited to intracellular localization signals.

A "chloroplast transit peptide" is an amino acid sequence which is translated in conjunction with a protein and directs the protein to the chloroplast or other plastid types present in the cell in which the protein is made. "Chloroplast transit sequence" refers to a nucleotide sequence that encodes a chloroplast transit peptide. A "signal peptide" is an amino acid sequence which is translated in conjunction with a protein and directs the protein to the secretory system (Chrispeels (1991) *Ann. Rev. Plant Phys. Plant Mol. Biol. 42*:21-53). If the protein is to be directed to a vacuole, a vacuolar targeting signal (*supra*) can further be added, or if to the endoplasmic reticulum, an endoplasmic reticulum retention signal (*supra*) may be added. If the protein is to be directed to the nucleus, any signal peptide present should be removed and instead a nuclear localization signal included (Raikhel (1992) *Plant Phys. 100*:1627-1632).

"Transformation" refers to the transfer of a nucleic acid fragment into the genome of a host organism, resulting in genetically stable inheritance. Host organisms containing the transformed nucleic acid fragments are referred to as "transgenic" organisms. Examples of methods of plant transformation include *Agrobacterium*-mediated transformation (De Blaere et al. (1987) *Meth. Enzymol. 143*:277) and particle-accelerated or "gene gun" transformation technology (Klein et al. (1987) *Nature (London) 327*:70-73; U.S. Patent No. 4,945,050, incorporated herein by reference).

Standard recombinant DNA and molecular cloning techniques used herein are well known in the art and are described more fully in Sambrook et al. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, 1989 (hereinafter "Maniatis").

Nucleic acid fragments encoding at least a portion of several flavonoid biosynthetic enzymes have been isolated and identified by comparison of random plant cDNA sequences to public databases containing nucleotide and protein sequences using the BLAST

algorithms well known to those skilled in the art. The nucleic acid fragments of the instant invention may be used to isolate cDNAs and genes encoding homologous proteins from the same or other plant species. Isolation of homologous genes using sequence-dependent protocols is well known in the art. Examples of sequence-dependent protocols include, but are not limited to, methods of nucleic acid hybridization, and methods of DNA and RNA amplification as exemplified by various uses of nucleic acid amplification technologies (e.g., polymerase chain reaction, ligase chain reaction).

For example, genes encoding other isoflavone 2-hydroxylase, either as cDNAs or genomic DNAs, could be isolated directly by using all or a portion of the instant nucleic acid fragments as DNA hybridization probes to screen libraries from any desired plant employing methodology well known to those skilled in the art. Specific oligonucleotide probes based upon the instant nucleic acid sequences can be designed and synthesized by methods known in the art (Maniatis). Moreover, the entire sequences can be used directly to synthesize DNA probes by methods known to the skilled artisan such as random primer DNA labeling, nick translation, or end-labeling techniques, or RNA probes using available *in vitro* transcription systems. In addition, specific primers can be designed and used to amplify a part or all of the instant sequences. The resulting amplification products can be labeled directly during amplification reactions or labeled after amplification reactions, and used as probes to isolate full length cDNA or genomic fragments under conditions of appropriate stringency.

In addition, two short segments of the instant nucleic acid fragments may be used in polymerase chain reaction protocols to amplify longer nucleic acid fragments encoding homologous genes from DNA or RNA. The polymerase chain reaction may also be performed on a library of cloned nucleic acid fragments wherein the sequence of one primer is derived from the instant nucleic acid fragments, and the sequence of the other primer takes advantage of the presence of the polyadenylic acid tracts to the 3' end of the mRNA precursor encoding plant genes. Alternatively, the second primer sequence may be based upon sequences derived from the cloning vector. For example, the skilled artisan can follow the RACE protocol (Frohman et al. (1988) *Proc. Natl. Acad. Sci. USA 85*:8998-9002) to generate cDNAs by using PCR to amplify copies of the region between a single point in the transcript and the 3' or 5' end. Primers oriented in the 3' and 5' directions can be designed from the instant sequences. Using commercially available 3' RACE or 5' RACE systems (BRL), specific 3' or 5' cDNA fragments can be isolated (Ohara et al. (1989) *Proc. Natl. Acad. Sci. USA 86*:5673-5677; Loh et al. (1989) *Science 243*:217-220). Products generated by the 3' and 5' RACE procedures can be combined to generate full-length cDNAs (Frohman and Martin (1989) *Techniques 1*:165). Consequently, a polynucleotide comprising a nucleotide sequence of at least one of 60 (preferably one of at least 40, most preferably one of at least 30) contiguous nucleotides derived from a nucleotide sequence selected from the

group consisting of SEQ ID NOs:1, 3, 5 and the complement of such nucleotide sequences may be used in such methods to obtain a nucleic acid fragment encoding a substantial portion of an amino acid sequence of a polypeptide. The present invention relates to a method of obtaining a nucleic acid fragment encoding a substantial portion of a polypeptide of a gene (such as isoflavone 2-hydroxylase) preferably a substantial portion of a plant polypeptide of a gene, comprising the steps of: synthesizing an oligonucleotide primer comprising a nucleotide sequence of at least one of 60 (preferably at least one of 40, most preferably at least one of 30) contiguous nucleotides derived from a nucleotide sequence selected from the group consisting of SEQ ID NOs:1, 3, 5 and the complement of such nucleotide sequences; and amplifying a nucleic acid fragment (preferably a cDNA inserted in a cloning vector) using the oligonucleotide primer. The amplified nucleic acid fragment preferably will encode a portion of a polypeptide (such as isoflavone 2-hydroxylase).

Availability of the instant nucleotide and deduced amino acid sequences facilitates immunological screening of cDNA expression libraries. Synthetic peptides representing portions of the instant amino acid sequences may be synthesized. These peptides can be used to immunize animals to produce polyclonal or monoclonal antibodies with specificity for peptides or proteins comprising the amino acid sequences. These antibodies can be then be used to screen cDNA expression libraries to isolate full-length cDNA clones of interest (Lerner (1984) *Adv. Immunol. 36*:1-34; Maniatis).

The nucleic acid fragments of the instant invention may be used to create transgenic plants in which the disclosed polypeptides are present at higher or lower levels than normal or in cell types or developmental stages in which they are not normally found. This would have the effect of altering the level of isoflavone 2-hydroxylase activity in those cells.

Overexpression of the proteins of the instant invention may be accomplished by first constructing a chimeric gene in which the coding region is operably linked to a promoter capable of directing expression of a gene in the desired tissues at the desired stage of development. The chimeric gene may comprise promoter sequences and translation leader sequences derived from the same genes. 3' Non-coding sequences encoding transcription termination signals may also be provided. The instant chimeric gene may also comprise one or more introns in order to facilitate gene expression.

Plasmid vectors comprising the isolated polynucleotide (or chimeric gene) may be constructed. The choice of plasmid vector is dependent upon the method that will be used to transform host plants. The skilled artisan is well aware of the genetic elements that must be present on the plasmid vector in order to successfully transform, select and propagate host cells containing the chimeric gene. The skilled artisan will also recognize that different independent transformation events will result in different levels and patterns of expression (Jones et al. (1985) *EMBO J. 4*:2411-2418; De Almeida et al. (1989) *Mol. Gen. Genetics 218*:78-86), and thus that multiple events must be screened in order to obtain lines

displaying the desired expression level and pattern. Such screening may be accomplished by Southern analysis of DNA, Northern analysis of mRNA expression, Western analysis of protein expression, or phenotypic analysis.

For some applications it may be useful to direct the instant polypeptides to different cellular compartments, or to facilitate its secretion from the cell. It is thus envisioned that the chimeric gene described above may be further supplemented by directing the coding sequence to encode the instant polypeptides with appropriate intracellular targeting sequences such as transit sequences (Keegstra (1989) *Cell 56*:247-253), signal sequences or sequences encoding endoplasmic reticulum localization (Chrispeels (1991) *Ann. Rev. Plant Phys. Plant Mol. Biol. 42*:21-53), or nuclear localization signals (Raikhel (1992) *Plant Phys.100*:1627-1632) with or without removing targeting sequences that are already present. While the references cited give examples of each of these, the list is not exhaustive and more targeting signals of use may be discovered in the future.

It may also be desirable to reduce or eliminate expression of genes encoding the instant polypeptides in plants for some applications. In order to accomplish this, a chimeric gene designed for co-suppression of the instant polypeptide can be constructed by linking a gene or gene fragment encoding that polypeptide to plant promoter sequences. Alternatively, a chimeric gene designed to express antisense RNA for all or part of the instant nucleic acid fragment can be constructed by linking the gene or gene fragment in reverse orientation to plant promoter sequences. Either the co-suppression or antisense chimeric genes could be introduced into plants via transformation wherein expression of the corresponding endogenous genes are reduced or eliminated.

Molecular genetic solutions to the generation of plants with altered gene expression have a decided advantage over more traditional plant breeding approaches. Changes in plant phenotypes can be produced by specifically inhibiting expression of one or more genes by antisense inhibition or cosuppression (U.S. Patent Nos. 5,190,931, 5,107,065 and 5,283,323). An antisense or cosuppression construct would act as a dominant negative regulator of gene activity. While conventional mutations can yield negative regulation of gene activity these effects are most likely recessive. The dominant negative regulation available with a transgenic approach may be advantageous from a breeding perspective. In addition, the ability to restrict the expression of specific phenotype to the reproductive tissues of the plant by the use of tissue specific promoters may confer agronomic advantages relative to conventional mutations which may have an effect in all tissues in which a mutant gene is ordinarily expressed.

The person skilled in the art will know that special considerations are associated with the use of antisense or cosuppression technologies in order to reduce expression of particular genes. For example, the proper level of expression of sense or antisense genes may require the use of different chimeric genes utilizing different regulatory elements known to the

skilled artisan. Once transgenic plants are obtained by one of the methods described above, it will be necessary to screen individual transgenics for those that most effectively display the desired phenotype. Accordingly, the skilled artisan will develop methods for screening large numbers of transformants. The nature of these screens will generally be chosen on practical grounds. For example, one can screen by looking for changes in gene expression by using antibodies specific for the protein encoded by the gene being suppressed, or one could establish assays that specifically measure enzyme activity. A preferred method will be one which allows large numbers of samples to be processed rapidly, since it will be expected that a large number of transformants will be negative for the desired phenotype.

The instant polypeptides (or portions thereof) may be produced in heterologous host cells, particularly in the cells of microbial hosts, and can be used to prepare antibodies to the these proteins by methods well known to those skilled in the art. The antibodies are useful for detecting the polypeptides of the instant invention *in situ* in cells or *in vitro* in cell extracts. Preferred heterologous host cells for production of the instant polypeptides are microbial hosts. Microbial expression systems and expression vectors containing regulatory sequences that direct high level expression of foreign proteins are well known to those skilled in the art. Any of these could be used to construct a chimeric gene for production of the instant polypeptides. This chimeric gene could then be introduced into appropriate microorganisms via transformation to provide high level expression of the encoded flavonoid biosynthetic enzyme. An example of a vector for high level expression of the instant polypeptides in a bacterial host is provided (Example 6).

All or a substantial portion of the nucleic acid fragments of the instant invention may also be used as probes for genetically and physically mapping the genes that they are a part of, and as markers for traits linked to those genes. Such information may be useful in plant breeding in order to develop lines with desired phenotypes. For example, the instant nucleic acid fragments may be used as restriction fragment length polymorphism (RFLP) markers. Southern blots (Maniatis) of restriction-digested plant genomic DNA may be probed with the nucleic acid fragments of the instant invention. The resulting banding patterns may then be subjected to genetic analyses using computer programs such as MapMaker (Lander et al. (1987) *Genomics 1*:174-181) in order to construct a genetic map. In addition, the nucleic acid fragments of the instant invention may be used to probe Southern blots containing restriction endonuclease-treated genomic DNAs of a set of individuals representing parent and progeny of a defined genetic cross. Segregation of the DNA polymorphisms is noted and used to calculate the position of the instant nucleic acid sequence in the genetic map previously obtained using this population (Botstein et al. (1980) *Am. J. Hum. Genet. 32*:314-331).

The production and use of plant gene-derived probes for use in genetic mapping is described in Bernatzky and Tanksley (1986) *Plant Mol. Biol. Reporter 4*:37-41. Numerous

publications describe genetic mapping of specific cDNA clones using the methodology outlined above or variations thereof. For example, F2 intercross populations, backcross populations, randomly mated populations, near isogenic lines, and other sets of individuals may be used for mapping. Such methodologies are well known to those skilled in the art.

Nucleic acid probes derived from the instant nucleic acid sequences may also be used for physical mapping (i.e., placement of sequences on physical maps; *see* Hoheisel et al. In: *Nonmammalian Genomic Analysis: A Practical Guide*, Academic press 1996, pp. 319-346, and references cited therein).

In another embodiment, nucleic acid probes derived from the instant nucleic acid sequences may be used in direct fluorescence *in situ* hybridization (FISH) mapping (Trask (1991) *Trends Genet. 7*:149-154). Although current methods of FISH mapping favor use of large clones (several to several hundred KB; see Laan et al. (1995) *Genome Res. 5*:13-20), improvements in sensitivity may allow performance of FISH mapping using shorter probes.

A variety of nucleic acid amplification-based methods of genetic and physical mapping may be carried out using the instant nucleic acid sequences. Examples include allele-specific amplification (Kazazian (1989) *J. Lab. Clin. Med. 11*:95-96), polymorphism of PCR-amplified fragments (CAPS; Sheffield et al. (1993) *Genomics 16*:325-332), allele-specific ligation (Landegren et al. (1988) *Science 241*:1077-1080), nucleotide extension reactions (Sokolov (1990) *Nucleic Acid Res. 18*:3671), Radiation Hybrid Mapping (Walter et al. (1997) *Nat. Genet. 7*:22-28) and Happy Mapping (Dear and Cook (1989) *Nucleic Acid Res. 17*:6795-6807). For these methods, the sequence of a nucleic acid fragment is used to design and produce primer pairs for use in the amplification reaction or in primer extension reactions. The design of such primers is well known to those skilled in the art. In methods employing PCR-based genetic mapping, it may be necessary to identify DNA sequence differences between the parents of the mapping cross in the region corresponding to the instant nucleic acid sequence. This, however, is generally not necessary for mapping methods.

Loss of function mutant phenotypes may be identified for the instant cDNA clones either by targeted gene disruption protocols or by identifying specific mutants for these genes contained in a maize population carrying mutations in all possible genes (Ballinger and Benzer (1989) *Proc. Natl. Acad. Sci USA 86*:9402-9406; Koes et al. (1995) *Proc. Natl. Acad. Sci USA 92*:8149-8153; Bensen et al. (1995) *Plant Cell 7*:75-84). The latter approach may be accomplished in two ways. First, short segments of the instant nucleic acid fragments may be used in polymerase chain reaction protocols in conjunction with a mutation tag sequence primer on DNAs prepared from a population of plants in which Mutator transposons or some other mutation-causing DNA element has been introduced (see Bensen, *supra*). The amplification of a specific DNA fragment with these primers indicates the insertion of the mutation tag element in or near the plant gene encoding the

instant polypeptides. Alternatively, the instant nucleic acid fragment may be used as a hybridization probe against PCR amplification products generated from the mutation population using the mutation tag sequence primer in conjunction with an arbitrary genomic site primer, such as that for a restriction enzyme site-anchored synthetic adaptor. With either method, a plant containing a mutation in the endogenous gene encoding the instant polypeptides can be identified and obtained. This mutant plant can then be used to determine or confirm the natural function of the instant polypeptides disclosed herein.

EXAMPLES

The present invention is further defined in the following Examples, in which all parts and percentages are by weight and degrees are Celsius, unless otherwise stated. It should be understood that these Examples, while indicating preferred embodiments of the invention, are given by way of illustration only. From the above discussion and these Examples, one skilled in the art can ascertain the essential characteristics of this invention, and without departing from the spirit and scope thereof, can make various changes and modifications of the invention to adapt it to various usages and conditions.

EXAMPLE 1

Composition of cDNA Libraries; Isolation and Sequencing of cDNA Clones

cDNA libraries representing mRNAs from various soybean tissues were prepared. The characteristics of the libraries are described below.

TABLE 2

cDNA Libraries from Soybean

Library	Tissue	Clone
sls1c	Soybean Infected With *Sclerotinia sclerotiorum* Mycelium	sls1c.pk005.n3
src3c	Soybean 8 Day Old Root Infected With Cyst Nematode *Heterodera glycinesis*	src3c.pk005.f5
sgc1c	Soybean Seeds 4 Hours After Germination	sgc1c.pk001.g17
sgs2c	Soybean Seeds 14 Hours After Germination	sgs2c.pk004.h7
sfl1	Soybean Immature Flower	sfl1.pk0034.g1

cDNA libraries may be prepared by any one of many methods available. For example, the cDNAs may be introduced into plasmid vectors by first preparing the cDNA libraries in Uni-ZAP™ XR vectors according to the manufacturer's protocol (Stratagene Cloning Systems, La Jolla, CA). The Uni-ZAP™ XR libraries are converted into plasmid libraries according to the protocol provided by Stratagene. Upon conversion, cDNA inserts will be contained in the plasmid vector pBluescript. In addition, the cDNAs may be introduced directly into precut Bluescript II SK(+) vectors (Stratagene) using T4 DNA ligase (New England Biolabs), followed by transfection into DH10B cells according to the

manufacturer's protocol (GIBCO BRL Products). Once the cDNA inserts are in plasmid vectors, plasmid DNAs are prepared from randomly picked bacterial colonies containing recombinant pBluescript plasmids, or the insert cDNA sequences are amplified via polymerase chain reaction using primers specific for vector sequences flanking the inserted cDNA sequences. Amplified insert DNAs or plasmid DNAs are sequenced in dye-primer sequencing reactions to generate partial cDNA sequences (expressed sequence tags or "ESTs"; see Adams et al., (1991) *Science 252*:1651-1656). The resulting ESTs are analyzed using a Perkin Elmer Model 377 fluorescent sequencer.

EXAMPLE 2

Identification of cDNA Clones

cDNA clones encoding flavonoid biosynthetic enzymes were identified by conducting BLAST (Basic Local Alignment Search Tool; Altschul et al. (1993) *J. Mol. Biol. 215*:403-410; see also www.ncbi.nlm.nih.gov/BLAST/) searches for similarity to sequences contained in the BLAST "nr" database (comprising all non-redundant GenBank CDS translations, sequences derived from the 3-dimensional structure Brookhaven Protein Data Bank, the last major release of the SWISS-PROT protein sequence database, EMBL, and DDBJ databases). The cDNA sequences obtained in Example 1 were analyzed for similarity to all publicly available DNA sequences contained in the "nr" database using the BLASTN algorithm provided by the National Center for Biotechnology Information (NCBI). The DNA sequences were translated in all reading frames and compared for similarity to all publicly available protein sequences contained in the "nr" database using the BLASTX algorithm (Gish and States (1993) *Nat. Genet. 3*:266-272) provided by the NCBI. For convenience, the P-value (probability) of observing a match of a cDNA sequence to a sequence contained in the searched databases merely by chance as calculated by BLAST are reported herein as "pLog" values, which represent the negative of the logarithm of the reported P-value. Accordingly, the greater the pLog value, the greater the likelihood that the cDNA sequence and the BLAST "hit" represent homologous proteins.

EXAMPLE 3

Characterization of cDNA Clones Encoding Isoflavone 2-Hydroxylase

The BLASTX search using the EST sequences from clones listed in Table 3 revealed similarity of the polypeptides encoded by the cDNAs to isoflavone 2-hydroxylase from *Glycyrrhiza echinta* (NCBI Identifier No. gi 4200044) and *Cicer arietinum* (NCBI Identifier No. gi 3850630). Shown in Table 3 are the BLAST results for individual ESTs ("EST"), the sequences of the entire cDNA inserts comprising the indicated cDNA clones ("FIS"), contigs assembled from two or more ESTs ("Contig"), contigs assembled from an FIS and one or more ESTs ("Contig*"), or sequences encoding the entire protein derived from an FIS, a contig, or an FIS and PCR ("CGS"):

TABLE 3

BLAST Results for Sequences Encoding Polypeptides Homologous to *Glycyrrhiza echinta* and *Cicer arietinum* Isoflavone 2-Hydroxylase

Clone	Status	BLAST pLog Score
sls1c.pk005.n3	(FIS)	>254.00 (gi 4200044)
src3c.pk005.f5	(FIS)	>254.00 (gi 3850630)

The data in Table 4 represents a calculation of the percent identity of the amino acid sequences set forth in SEQ ID NOs:2 and 4 and the *Glycyrrhiza echinta* and *Cicer arietinum* sequences.

TABLE 4

Percent Identity of Amino Acid Sequences Deduced From the Nucleotide Sequences of cDNA Clones Encoding Polypeptides Homologous to *Glycyrrhiza echinta* and *Cicer arietinum* Isoflavone 2-Hydroxylase

SEQ ID NO.	Percent Identity to
2	60% (gi 4200044)
4	61% (gi 3850630)

Sequence alignments and percent identity calculations were performed using the Megalign program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). Multiple alignment of the sequences was performed using the Clustal method of alignment (Higgins and Sharp (1989) *CABIOS. 5*:151-153) with the default parameters (GAP PENALTY=10, GAP LENGTH PENALTY=10). Default parameters for pairwise alignments using the Clustal method were KTUPLE 1, GAP PENALTY=3, WINDOW=5 and DIAGONALS SAVED=5. Sequence alignments and BLAST scores and probabilities indicate that the nucleic acid fragments comprising the instant cDNA clones encode a substantial portion of a isoflavone 2-hydroxylase. These sequences represent the first soybean sequences encoding isoflavone 2-hydroxylase.

EXAMPLE 4

Expression of Chimeric Genes in Monocot Cells

A chimeric gene comprising a cDNA encoding the instant polypeptides in sense orientation with respect to the maize 27 kD zein promoter that is located 5' to the cDNA fragment, and the 10 kD zein 3' end that is located 3' to the cDNA fragment, can be constructed. The cDNA fragment of this gene may be generated by polymerase chain reaction (PCR) of the cDNA clone using appropriate oligonucleotide primers. Cloning sites (NcoI or SmaI) can be incorporated into the oligonucleotides to provide proper orientation of the DNA fragment when inserted into the digested vector pML103 as described below. Amplification is then performed in a standard PCR. The amplified DNA is then digested

with restriction enzymes NcoI and SmaI and fractionated on an agarose gel. The appropriate band can be isolated from the gel and combined with a 4.9 kb NcoI-SmaI fragment of the plasmid pML103. Plasmid pML103 has been deposited under the terms of the Budapest Treaty at ATCC (American Type Culture Collection, 10801 University Blvd., Manassas, VA 20110-2209), and bears accession number ATCC 97366. The DNA segment from pML103 contains a 1.05 kb SalI-NcoI promoter fragment of the maize 27 kD zein gene and a 0.96 kb SmaI-SalI fragment from the 3' end of the maize 10 kD zein gene in the vector pGem9Zf(+) (Promega). Vector and insert DNA can be ligated at 15°C overnight, essentially as described (Maniatis). The ligated DNA may then be used to transform *E. coli* XL1-Blue (Epicurian Coli XL-1 Blue™; Stratagene). Bacterial transformants can be screened by restriction enzyme digestion of plasmid DNA and limited nucleotide sequence analysis using the dideoxy chain termination method (Sequenase™ DNA Sequencing Kit; U.S. Biochemical). The resulting plasmid construct would comprise a chimeric gene encoding, in the 5' to 3' direction, the maize 27 kD zein promoter, a cDNA fragment encoding the instant polypeptides, and the 10 kD zein 3' region.

The chimeric gene described above can then be introduced into corn cells by the following procedure. Immature corn embryos can be dissected from developing caryopses derived from crosses of the inbred corn lines H99 and LH132. The embryos are isolated 10 to 11 days after pollination when they are 1.0 to 1.5 mm long. The embryos are then placed with the axis-side facing down and in contact with agarose-solidified N6 medium (Chu et al. (1975) *Sci. Sin. Peking* 18:659-668). The embryos are kept in the dark at 27°C. Friable embryogenic callus consisting of undifferentiated masses of cells with somatic proembryoids and embryoids borne on suspensor structures proliferates from the scutellum of these immature embryos. The embryogenic callus isolated from the primary explant can be cultured on N6 medium and sub-cultured on this medium every 2 to 3 weeks.

The plasmid, p35S/Ac (obtained from Dr. Peter Eckes, Hoechst Ag, Frankfurt, Germany) may be used in transformation experiments in order to provide for a selectable marker. This plasmid contains the *Pat* gene (see European Patent Publication 0 242 236) which encodes phosphinothricin acetyl transferase (PAT). The enzyme PAT confers resistance to herbicidal glutamine synthetase inhibitors such as phosphinothricin. The *pat* gene in p35S/Ac is under the control of the 35S promoter from Cauliflower Mosaic Virus (Odell et al. (1985) *Nature* 313:810-812) and the 3' region of the nopaline synthase gene from the T-DNA of the Ti plasmid of *Agrobacterium tumefaciens*.

The particle bombardment method (Klein et al. (1987) *Nature* 327:70-73) may be used to transfer genes to the callus culture cells. According to this method, gold particles (1 μm in diameter) are coated with DNA using the following technique. Ten μg of plasmid DNAs are added to 50 μL of a suspension of gold particles (60 mg per mL). Calcium chloride (50 μL of a 2.5 M solution) and spermidine free base (20 μL of a 1.0 M solution) are added

to the particles. The suspension is vortexed during the addition of these solutions. After 10 minutes, the tubes are briefly centrifuged (5 sec at 15,000 rpm) and the supernatant removed. The particles are resuspended in 200 μL of absolute ethanol, centrifuged again and the supernatant removed. The ethanol rinse is performed again and the particles resuspended in a final volume of 30 μL of ethanol. An aliquot (5 μL) of the DNA-coated gold particles can be placed in the center of a Kapton™ flying disc (Bio-Rad Labs). The particles are then accelerated into the corn tissue with a Biolistic™ PDS-1000/He (Bio-Rad Instruments, Hercules CA), using a helium pressure of 1000 psi, a gap distance of 0.5 cm and a flying distance of 1.0 cm.

For bombardment, the embryogenic tissue is placed on filter paper over agarose-solidified N6 medium. The tissue is arranged as a thin lawn and covered a circular area of about 5 cm in diameter. The petri dish containing the tissue can be placed in the chamber of the PDS-1000/He approximately 8 cm from the stopping screen. The air in the chamber is then evacuated to a vacuum of 28 inches of Hg. The macrocarrier is accelerated with a helium shock wave using a rupture membrane that bursts when the He pressure in the shock tube reaches 1000 psi.

Seven days after bombardment the tissue can be transferred to N6 medium that contains gluphosinate (2 mg per liter) and lacks casein or proline. The tissue continues to grow slowly on this medium. After an additional 2 weeks the tissue can be transferred to fresh N6 medium containing gluphosinate. After 6 weeks, areas of about 1 cm in diameter of actively growing callus can be identified on some of the plates containing the glufosinate-supplemented medium. These calli may continue to grow when sub-cultured on the selective medium.

Plants can be regenerated from the transgenic callus by first transferring clusters of tissue to N6 medium supplemented with 0.2 mg per liter of 2,4-D. After two weeks the tissue can be transferred to regeneration medium (Fromm et al. (1990) *Bio/Technology 8*:833-839).

EXAMPLE 5

Expression of Chimeric Genes in Dicot Cells

A seed-specific expression cassette composed of the promoter and transcription terminator from the gene encoding the β subunit of the seed storage protein phaseolin from the bean *Phaseolus vulgaris* (Doyle et al. (1986) *J. Biol. Chem. 261*:9228-9238) can be used for expression of the instant polypeptides in transformed soybean. The phaseolin cassette includes about 500 nucleotides upstream (5') from the translation initiation codon and about 1650 nucleotides downstream (3') from the translation stop codon of phaseolin. Between the 5' and 3' regions are the unique restriction endonuclease sites Nco I (which includes the ATG translation initiation codon), Sma I, Kpn I and Xba I. The entire cassette is flanked by Hind III sites.

The cDNA fragment of this gene may be generated by polymerase chain reaction (PCR) of the cDNA clone using appropriate oligonucleotide primers. Cloning sites can be incorporated into the oligonucleotides to provide proper orientation of the DNA fragment when inserted into the expression vector. Amplification is then performed as described above, and the isolated fragment is inserted into a pUC18 vector carrying the seed expression cassette.

Soybean embryos may then be transformed with the expression vector comprising sequences encoding the instant polypeptides. To induce somatic embryos, cotyledons, 3-5 mm in length dissected from surface sterilized, immature seeds of the soybean cultivar A2872, can be cultured in the light or dark at 26°C on an appropriate agar medium for 6-10 weeks. Somatic embryos which produce secondary embryos are then excised and placed into a suitable liquid medium. After repeated selection for clusters of somatic embryos which multiplied as early, globular staged embryos, the suspensions are maintained as described below.

Soybean embryogenic suspension cultures can maintained in 35 mL liquid media on a rotary shaker, 150 rpm, at 26°C with florescent lights on a 16:8 hour day/night schedule. Cultures are subcultured every two weeks by inoculating approximately 35 mg of tissue into 35 mL of liquid medium.

Soybean embryogenic suspension cultures may then be transformed by the method of particle gun bombardment (Klein et al. (1987) *Nature* (London) *327*:70-73, U.S. Patent No. 4,945,050). A DuPont Biolistic™ PDS1000/HE instrument (helium retrofit) can be used for these transformations.

A selectable marker gene which can be used to facilitate soybean transformation is a chimeric gene composed of the 35S promoter from Cauliflower Mosaic Virus (Odell et al. (1985) *Nature 313*:810-812), the hygromycin phosphotransferase gene from plasmid pJR225 (from *E. coli*; Gritz et al.(1983) *Gene 25*:179-188) and the 3' region of the nopaline synthase gene from the T-DNA of the Ti plasmid of *Agrobacterium tumefaciens*. The seed expression cassette comprising the phaseolin 5' region, the fragment encoding the instant polypeptides and the phaseolin 3' region can be isolated as a restriction fragment. This fragment can then be inserted into a unique restriction site of the vector carrying the marker gene.

To 50 μL of a 60 mg/mL 1 μm gold particle suspension is added (in order): 5 μL DNA (1 μg/μL), 20 μl spermidine (0.1 M), and 50 μL $CaCl_2$ (2.5 M). The particle preparation is then agitated for three minutes, spun in a microfuge for 10 seconds and the supernatant removed. The DNA-coated particles are then washed once in 400 μL 70% ethanol and resuspended in 40 μL of anhydrous ethanol. The DNA/particle suspension can be sonicated three times for one second each. Five μL of the DNA-coated gold particles are then loaded on each macro carrier disk.

Approximately 300-400 mg of a two-week-old suspension culture is placed in an empty 60x15 mm petri dish and the residual liquid removed from the tissue with a pipette. For each transformation experiment, approximately 5-10 plates of tissue are normally bombarded. Membrane rupture pressure is set at 1100 psi and the chamber is evacuated to a vacuum of 28 inches mercury. The tissue is placed approximately 3.5 inches away from the retaining screen and bombarded three times. Following bombardment, the tissue can be divided in half and placed back into liquid and cultured as described above.

Five to seven days post bombardment, the liquid media may be exchanged with fresh media, and eleven to twelve days post bombardment with fresh media containing 50 mg/mL hygromycin. This selective media can be refreshed weekly. Seven to eight weeks post bombardment, green, transformed tissue may be observed growing from untransformed, necrotic embryogenic clusters. Isolated green tissue is removed and inoculated into individual flasks to generate new, clonally propagated, transformed embryogenic suspension cultures. Each new line may be treated as an independent transformation event. These suspensions can then be subcultured and maintained as clusters of immature embryos or regenerated into whole plants by maturation and germination of individual somatic embryos.

EXAMPLE 6

Expression of Chimeric Genes in Microbial Cells

The cDNAs encoding the instant polypeptides can be inserted into the T7 *E. coli* expression vector pBT430. This vector is a derivative of pET-3a (Rosenberg et al. (1987) *Gene 56*:125-135) which employs the bacteriophage T7 RNA polymerase/T7 promoter system. Plasmid pBT430 was constructed by first destroying the EcoR I and Hind III sites in pET-3a at their original positions. An oligonucleotide adaptor containing EcoR I and Hind III sites was inserted at the BamH I site of pET-3a. This created pET-3aM with additional unique cloning sites for insertion of genes into the expression vector. Then, the Nde I site at the position of translation initiation was converted to an Nco I site using oligonucleotide-directed mutagenesis. The DNA sequence of pET-3aM in this region, 5'-CATATGG, was converted to 5'-CCCATGG in pBT430.

Plasmid DNA containing a cDNA may be appropriately digested to release a nucleic acid fragment encoding the protein. This fragment may then be purified on a 1% NuSieve GTG™ low melting agarose gel (FMC). Buffer and agarose contain 10 μg/ml ethidium bromide for visualization of the DNA fragment. The fragment can then be purified from the agarose gel by digestion with GELase™ (Epicentre Technologies) according to the manufacturer's instructions, ethanol precipitated, dried and resuspended in 20 μL of water. Appropriate oligonucleotide adapters may be ligated to the fragment using T4 DNA ligase (New England Biolabs, Beverly, MA). The fragment containing the ligated adapters can be purified from the excess adapters using low melting agarose as described above. The vector pBT430 is digested, dephosphorylated with alkaline phosphatase (NEB) and deproteinized

with phenol/chloroform as described above. The prepared vector pBT430 and fragment can then be ligated at 16°C for 15 hours followed by transformation into DH5 electrocompetent cells (GIBCO BRL). Transformants can be selected on agar plates containing LB media and 100 μg/mL ampicillin. Transformants containing the gene encoding the instant polypeptides are then screened for the correct orientation with respect to the T7 promoter by restriction enzyme analysis.

For high level expression, a plasmid clone with the cDNA insert in the correct orientation relative to the T7 promoter can be transformed into *E. coli* strain BL21(DE3) (Studier et al. (1986) *J. Mol. Biol. 189*:113-130). Cultures are grown in LB medium containing ampicillin (100 mg/L) at 25°C. At an optical density at 600 nm of approximately 1, IPTG (isopropylthio-β-galactoside, the inducer) can be added to a final concentration of 0.4 mM and incubation can be continued for 3 h at 25°. Cells are then harvested by centrifugation and re-suspended in 50 μL of 50 mM Tris-HCl at pH 8.0 containing 0.1 mM DTT and 0.2 mM phenyl methylsulfonyl fluoride. A small amount of 1 mm glass beads can be added and the mixture sonicated 3 times for about 5 seconds each time with a microprobe sonicator. The mixture is centrifuged and the protein concentration of the supernatant determined. One μg of protein from the soluble fraction of the culture can be separated by SDS-polyacrylamide gel electrophoresis. Gels can be observed for protein bands migrating at the expected molecular weight.

Various modifications of the invention in addition to those shown and described herein will be apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims.

The disclosure of each reference set forth above is incorporated herein by reference in its entirety.

CLAIMS

What is claimed is:

1. An isolated polynucleotide comprising a first nucleotide sequence encoding a polypeptide of at least 494 amino acids that has at least 80 % identity based on the Clustal method of alignment when compared to a polypeptide selected from the group consisting of SEQ ID NOs:2 and 4 or a second nucleotide sequence comprising the complement of the first nucleotide sequence.

2. The isolated polynucleotide of Claim 1, wherein the first nucleotide sequence consists of a nucleic acid sequence selected from the group consisting of SEQ ID NOs:1 and 3 that codes for the polypeptide selected from the group consisting of SEQ ID NOs:2 and 4.

3. The isolated polynucleotide of Claim 1 wherein the nucleotide sequences are DNA.

4. The isolated polynucleotide of Claim 1 wherein the nucleotide sequences are RNA.

5. A chimeric gene comprising the isolated polynucleotide of Claim 1 operably linked to suitable regulatory sequences.

6. A host cell comprising the chimeric gene of Claim 5.

7. A host cell comprising an isolated polynucleotide of Claim 1.

8. The host cell of Claim 7 wherein the host cell is selected from the group consisting of yeast, bacteria, plant, and virus.

9. A virus comprising the isolated polynucleotide of Claim 1.

10. A polypeptide of at least 494 amino acids that has at least 80% identity based on the Clustal method of alignment when compared to a polypeptide selected from the group consisting of SEQ ID NOs:2 and 4.

11. A method of selecting an isolated polynucleotide that affects the level of expression of a flavonoid biosynthetic enzyme polypeptide in a plant cell, the method comprising the steps of:

(a) constructing an isolated polynucleotide comprising a nucleotide sequence of at least one of 30 contiguous nucleotides derived from an isolated polynucleotide of Claim 1;

(b) introducing the isolated polynucleotide into a plant cell;

(c) measuring the level of a polypeptide in the plant cell containing the polynucleotide; and

(d) comparing the level of polypeptide in the plant cell containing the isolated polynucleotide with the level of polypeptide in a plant cell that does not contain the isolated polynucleotide.

12. The method of Claim 11 wherein the isolated polynucleotide consists of a nucleotide sequence selected from the group consisting of SEQ ID NOs:1 and 3 that codes for the polypeptide selected from the group consisting of SEQ ID NOs:2 and 4.

13. A method of selecting an isolated polynucleotide that affects the level of expression of a flavonoid biosynthetic enzyme polypeptide in a plant cell, the method comprising the steps of:

(a) constructing an isolated polynucleotide of Claim 1;

(b) introducing the isolated polynucleotide into a plant cell; and

(c) measuring the level of polypeptide in the plant cell containing the polynucleotide to provide a positive selection means.

14. A method of obtaining a nucleic acid fragment encoding a flavonoid biosynthetic enzyme polypeptide comprising the steps of:

(a) synthesizing an oligonucleotide primer comprising a nucleotide sequence of at least one of 30 contiguous nucleotides derived from a nucleotide sequence selected from the group consisting of SEQ ID NOs:1 and 3and the complement of such nucleotide sequences; and

(b) amplifying a nucleic acid sequence using the oligonucleotide primer.

15. A method of obtaining a nucleic acid fragment encoding a flavonoid biosynthetic enzyme polypeptide comprising the steps of:

(a) probing a cDNA or genomic library with an isolated polynucleotide comprising at least one of 30 contiguous nucleotides derived from a nucleotide sequence selected from the group consisting of SEQ ID NOs:1, 3 and the complement of such nucleotide sequences;

(b) identifying a DNA clone that hybridizes with the isolated polynucleotide;

(c) isolating the identified DNA clone; and

(d) sequencing the cDNA or genomic fragment that comprises the isolated DNA clone.

16. A composition comprising the isolated polynucleotide of Claim 1.

17. A composition comprising the isolated polypeptide of Claim 10.

18. An isolated polynucleotide comprising the nucleotide sequence having at least one of 30 contiguous nucleotides derived from a nucleic acid sequence selected from the group consisting of SEQ ID NOs:1, 3 and the complement of such sequences.

19. An expression cassette comprising an isolated polynucleotide of Claim 1 operably linked to a promoter.

20. A method for positive selection of a transformed cell comprising:

(a) transforming a host cell with the chimeric gene of Claim 5 or an expression cassette of Claim 19; and

(b) growing the transformed host cell under conditions which allow expression of the polynucleotide in an amount sufficient to complement a mutant cell with altered isoflavone 2-hydroxylase activity to provide a positive selection means.

21. The method of any one of Claims 11 or 13 wherein the plant cell is a monocot.

22. The method of any one of Claims 11 or 13 wherein the plant cell is a dicot.

23. An isolated polynucleotide comprising a first nucleotide sequence encoding a polypeptide of at least 141 amino acids that has at least 80 % identity based on the Clustal method of alignment when compared to a polypeptide of SEQ ID NO:6 or a second nucleotide sequence comprising the complement of the first nucleotide sequence.

24. A polypeptide comprising at least 141 amino acids that has at least 80% identity based on the Clustal method of alignment when compared to a polypeptide of SEQ ID NO:6.

SEQUENCE LISTING

<110> E. I. du Pont de Nemours and Company

<120> Flavonoid Biosynthetic Enzymes

<130> BB1324

<140>
<141>

<150> 60/113,190
<151> 1998-12-21

<160> 6

<170> Microsoft Office 97

<210> 1
<211> 1859
<212> DNA
<213> Glycine max

<400> 1
gaaaacactg acagacagca tagtctctgg tgcaagaatc aattgagcaa gcatgggaat 60
gttgttggtg gtggtctcat acgctgtcct tttcctggtt ctattcctcg gcgtgaagtt 120
tgttttccaa agcagaaaat tgagaaacat accaccaggt cctcctcctc ttcccataat 180
aggaaacctt aacctcctcg aacagccaat ccaccgtttc ttccaacgca tgtcgaaaca 240
gtacggcaac gtggtttccc tctggttcgg ttcacgtctg gccgttgtca tctcctctcc 300
aacagcatac caagaatgct tcaccaaaca cgacgttgcc ttggccaacc ggctaccttc 360
tctctcggga aaatacatct tctacaacaa caccaccgta ggctcctgct cccacggcga 420
gcactggcgc aacctccgcc gcatcaccgc cctggacgtc ctctccacgc agcgcgtcca 480
ctccttctcc ggaatccgga gcgacgagac gaagcgtctg atgcagaggt tggtgctggc 540
caagaactcg aacgaggaag agtttgcgcg agtggagatt agttcgatgt tcaacgactt 600
aacttacaac aacataatga ggatgatatc ggggaagagg ttttacggag aggagagtga 660
gatgaagaac gttgaggaag cgagggagtt cagagagact gtgacagaaa tgttggaact 720
catgggcttg gctaacaagg gagatcactt gcctttcctc aggtggttcg attttcagaa 780
tgtggagaag cgcttaaaga gtatcagtaa gaggtacgat tccatcttga ataagatcct 840
tcatgagaac cgtgccagca atgaccgcca gaattccatg atcgatcatc tcctcaaact 900
gcaagagacc cagcctcagt actacactga ccaaatcatc aaaggccttg ctctggccat 960
gctttttggt ggaactgact catcaactgg gactttagag tggtcattat ctaatttatt 1020
gaatcaccca gaggtgttga agaaggcaag agatgaattg gacactcaag tgggacaaga 1080
ccgcttgtta aatgagtcag accttccaaa acttccatat cttaggaaga tcatccttga 1140
gacacttagg ttgtaccccc cggccccaat tctaatacct catgtgtctt cagaagatat 1200
tacaattgaa ggattcaata tcccacgaga cacaattgtg atcattaatg gttggggcat 1260
gcagagagat cctcagttgt ggaatgatgc cacatgcttt aaacctgaga ggtttgatgt 1320
ggaaggagag gagaaaaagt tggtagcatt tggcatggga agaagggctt gcccaggaga 1380
acccatggct atgcaaagtg tcagctttac tttgggattg ttgattcaat gttttgactg 1440
gaaacgagta agtgaggaaa agcttgatat gacagagaac aattggatca ccttgtcaag 1500
gttaattcca ttggaggcca tgtgcaaggc tcgcccactt gccactaaaa ttggaattta 1560
attattaata gtatttttat ttggtaaact tgggtgattc agaatctaat acttataatt 1620
ttagtgtgta agagtggtga tcatatatac atttcaaaat taataatctt tgtccaaaaa 1680
tcatccatgg acaactatat gtcaattgac atctagagag aaatatagat ataagaatat 1740
ttatatttta ttactcttct ttatcttatg tgtcaaggcc cattgtagaa ttgggtgagc 1800
attaacatat atcaatattg tataccgccc agttttctca aataaatttc tttactttc 1859

<210> 2
<211> 499
<212> PRT
<213> Glycine max

<400> 2

```
Leu Leu Val Val Val Ser Tyr Ala Val Leu Phe Leu Val Leu Phe Leu
  1               5                  10                  15

Gly Val Lys Phe Val Phe Gln Ser Arg Lys Leu Arg Asn Ile Pro Pro
             20                  25                  30

Gly Pro Pro Pro Leu Pro Ile Ile Gly Asn Leu Asn Leu Leu Glu Gln
         35                  40                  45

Pro Ile His Arg Phe Phe Gln Arg Met Ser Lys Gln Tyr Gly Asn Val
     50                  55                  60

Val Ser Leu Trp Phe Gly Ser Arg Leu Ala Val Val Ile Ser Ser Pro
 65                  70                  75                  80

Thr Ala Tyr Gln Glu Cys Phe Thr Lys His Asp Val Ala Leu Ala Asn
                 85                  90                  95

Arg Leu Pro Ser Leu Ser Gly Lys Tyr Ile Phe Tyr Asn Asn Thr Thr
            100                 105                 110

Val Gly Ser Cys Ser His Gly Glu His Trp Arg Asn Leu Arg Arg Ile
        115                 120                 125

Thr Ala Leu Asp Val Leu Ser Thr Gln Arg Val His Ser Phe Ser Gly
    130                 135                 140

Ile Arg Ser Asp Glu Thr Lys Arg Leu Met Gln Arg Leu Val Leu Ala
145                 150                 155                 160

Lys Asn Ser Asn Glu Glu Glu Phe Ala Arg Val Glu Ile Ser Ser Met
                165                 170                 175

Phe Asn Asp Leu Thr Tyr Asn Asn Ile Met Arg Met Ile Ser Gly Lys
            180                 185                 190

Arg Phe Tyr Gly Glu Glu Ser Glu Met Lys Asn Val Glu Glu Ala Arg
        195                 200                 205

Glu Phe Arg Glu Thr Val Thr Glu Met Leu Glu Leu Met Gly Leu Ala
    210                 215                 220

Asn Lys Gly Asp His Leu Pro Phe Leu Arg Trp Phe Asp Phe Gln Asn
225                 230                 235                 240

Val Glu Lys Arg Leu Lys Ser Ile Ser Lys Arg Tyr Asp Ser Ile Leu
                245                 250                 255

Asn Lys Ile Leu His Glu Asn Arg Ala Ser Asn Asp Arg Gln Asn Ser
            260                 265                 270

Met Ile Asp His Leu Leu Lys Leu Gln Glu Thr Gln Pro Gln Tyr Tyr
        275                 280                 285

Thr Asp Gln Ile Ile Lys Gly Leu Ala Leu Ala Met Leu Phe Gly Gly
    290                 295                 300
```

```
Thr Asp Ser Ser Thr Gly Thr Leu Glu Trp Ser Leu Ser Asn Leu Leu
305                 310                 315                 320

Asn His Pro Glu Val Leu Lys Lys Ala Arg Asp Glu Leu Asp Thr Gln
                325                 330                 335

Val Gly Gln Asp Arg Leu Leu Asn Glu Ser Asp Leu Pro Lys Leu Pro
            340                 345                 350

Tyr Leu Arg Lys Ile Ile Leu Glu Thr Leu Arg Leu Tyr Pro Pro Ala
        355                 360                 365

Pro Ile Leu Ile Pro His Val Ser Ser Glu Asp Ile Thr Ile Glu Gly
    370                 375                 380

Phe Asn Ile Pro Arg Asp Thr Ile Val Ile Ile Asn Gly Trp Gly Met
385                 390                 395                 400

Gln Arg Asp Pro Gln Leu Trp Asn Asp Ala Thr Cys Phe Lys Pro Glu
                405                 410                 415

Arg Phe Asp Val Glu Gly Glu Glu Lys Lys Leu Val Ala Phe Gly Met
            420                 425                 430

Gly Arg Arg Ala Cys Pro Gly Glu Pro Met Ala Met Gln Ser Val Ser
        435                 440                 445

Phe Thr Leu Gly Leu Leu Ile Gln Cys Phe Asp Trp Lys Arg Val Ser
    450                 455                 460

Glu Glu Lys Leu Asp Met Thr Glu Asn Asn Trp Ile Thr Leu Ser Arg
465                 470                 475                 480

Leu Ile Pro Leu Glu Ala Met Cys Lys Ala Arg Pro Leu Ala Thr Lys
                485                 490                 495

Ile Gly Ile
```

<210> 3
<211> 1698
<212> DNA
<213> Glycine max

<400> 3

```
cagtaataac aatgtctcct ttcttatctt actctcttct ttccctcgtg ttcttcttca   60
ctctcaagta ccttttccaa agaagcagaa aagtacgaaa cctgccacct ggtccgactc  120
ctcttcctat aatcggcaac cttaacctcg ttgaacaacc tatacaccgt ttcttccacc  180
gcatgtccca aaaatatgga aacatcatat ccctttggtt tgggtcacgt cttgttgtgg  240
ttgtttcatc acccacagcg taccaagaat gtttcaccaa acatgatgtt accttggcca  300
acagggtacg ctccctctcg ggaaaataca tattctacga caacaccacc gtagggtctt  360
gctcccacgg cgagcactgg cgcaacctcc gccgcataac ctctctcgac gttctatcga  420
cgcagcgcgt ccactccttc tccggaatcc ggagcgacga gacgaagagg ttgatacaca  480
ggctggccag ggactccggg aaagattttg cgcgcgtgga gatgacctcc aagtttgctg  540
acttgacgta caacaacatc atgaggatga tttcggggaa gcggttttac ggagaagaga  600
gtgaacttaa caacgttgag gaagcgaagg agttcagaga cactgtgaat gagatgctgc  660
aactcatggg gttggctaac aagggagatc acttaccttt cctaaggtgg ttcgatttc   720
agaacgtgga gaagaggttg aagaatatca gtaagaggta tgataccatc ttgaataaga  780
tccttgatga gaaccgtaac aacaaggacc gcgagaattc catgattggt catctcctca  840
aactgcaaga gacacagcct gactattata ccgaccaaat catcaaaggc cttgctttgg  900
```

ctatgctctt tggtggaaca gactcgtcaa ctggaacttt agagtgggca ttatctaatt 960
tagtgaatga cccagaggtg ctgcagaagg caagagatga gttggacgct caagtaggac 1020
cagatcggct gttaaatgag tcagaccttc caaaacttcc ttatctcagg aagatagttc 1080
ttgaaacact taggttgtac cctccggctc caattctaat accacacgtg gcttcagaag 1140
acatcaatat cgaaggattc aatgttccac gagacacaat tgtgattatt aatggttggg 1200
ccatgcaaag agatcctaag atatggaaag atgcgacaag ctttaaacct gagaggtttg 1260
atgaagaagg agaggagaag aaattggtag catttggtat gggaagaagg gcttgcccag 1320
gagaacccat ggctatgcaa agtgttagct atactttggg attaatgatt caatgttttg 1380
actggaaacg agtaagtgag aagaagcttg atatgacaga gaataattgg atcaccttgt 1440
caaggttaat tccattggag gctatgtgta aagcccgccc actcgccagc aaagttgaaa 1500
gttattaaca atattttatt tggtatattt gggtgaggat ctaatactca taatttcggt 1560
gtgtaagtct atgcatgtta aaattaataa tatttgtcgt atgtccacaa ggccaaatgt 1620
agtactgggt gtggatttgc atatacaata tcaatattgt ataaatccca gtttccttga 1680
ataaatttct ttactttc 1698

<210> 4
<211> 494
<212> PRT
<213> Glycine max

<400> 4

```
Leu Ser Tyr Ser Leu Leu Ser Leu Val Phe Phe Phe Thr Leu Lys Tyr
  1               5                  10                  15

Leu Phe Gln Arg Ser Arg Lys Val Arg Asn Leu Pro Pro Gly Pro Thr
             20                  25                  30

Pro Leu Pro Ile Ile Gly Asn Leu Asn Leu Val Glu Gln Pro Ile His
         35                  40                  45

Arg Phe Phe His Arg Met Ser Gln Lys Tyr Gly Asn Ile Ile Ser Leu
     50                  55                  60

Trp Phe Gly Ser Arg Leu Val Val Val Val Ser Ser Pro Thr Ala Tyr
 65                  70                  75                  80

Gln Glu Cys Phe Thr Lys His Asp Val Thr Leu Ala Asn Arg Val Arg
                 85                  90                  95

Ser Leu Ser Gly Lys Tyr Ile Phe Tyr Asp Asn Thr Thr Val Gly Ser
            100                 105                 110

Cys Ser His Gly Glu His Trp Arg Asn Leu Arg Arg Ile Thr Ser Leu
        115                 120                 125

Asp Val Leu Ser Thr Gln Arg Val His Ser Phe Ser Gly Ile Arg Ser
    130                 135                 140

Asp Glu Thr Lys Arg Leu Ile His Arg Leu Ala Arg Asp Ser Gly Lys
145                 150                 155                 160

Asp Phe Ala Arg Val Glu Met Thr Ser Lys Phe Ala Asp Leu Thr Tyr
                165                 170                 175

Asn Asn Ile Met Arg Met Ile Ser Gly Lys Arg Phe Tyr Gly Glu Glu
            180                 185                 190

Ser Glu Leu Asn Asn Val Glu Glu Ala Lys Glu Phe Arg Asp Thr Val
        195                 200                 205
```

```
Asn Glu Met Leu Gln Leu Met Gly Leu Ala Asn Lys Gly Asp His Leu
    210                 215                 220

Pro Phe Leu Arg Trp Phe Asp Phe Gln Asn Val Glu Lys Arg Leu Lys
225                 230                 235                 240

Asn Ile Ser Lys Arg Tyr Asp Thr Ile Leu Asn Lys Ile Leu Asp Glu
                245                 250                 255

Asn Arg Asn Asn Lys Asp Arg Glu Asn Ser Met Ile Gly His Leu Leu
            260                 265                 270

Lys Leu Gln Glu Thr Gln Pro Asp Tyr Tyr Thr Asp Gln Ile Ile Lys
        275                 280                 285

Gly Leu Ala Leu Ala Met Leu Phe Gly Gly Thr Asp Ser Ser Thr Gly
    290                 295                 300

Thr Leu Glu Trp Ala Leu Ser Asn Leu Val Asn Asp Pro Glu Val Leu
305                 310                 315                 320

Gln Lys Ala Arg Asp Glu Leu Asp Ala Gln Val Gly Pro Asp Arg Leu
                325                 330                 335

Leu Asn Glu Ser Asp Leu Pro Lys Leu Pro Tyr Leu Arg Lys Ile Val
            340                 345                 350

Leu Glu Thr Leu Arg Leu Tyr Pro Pro Ala Pro Ile Leu Ile Pro His
        355                 360                 365

Val Ala Ser Glu Asp Ile Asn Ile Glu Gly Phe Asn Val Pro Arg Asp
    370                 375                 380

Thr Ile Val Ile Ile Asn Gly Trp Ala Met Gln Arg Asp Pro Lys Ile
385                 390                 395                 400

Trp Lys Asp Ala Thr Ser Phe Lys Pro Glu Arg Phe Asp Glu Glu Gly
                405                 410                 415

Glu Glu Lys Lys Leu Val Ala Phe Gly Met Gly Arg Arg Ala Cys Pro
            420                 425                 430

Gly Glu Pro Met Ala Met Gln Ser Val Ser Tyr Thr Leu Gly Leu Met
        435                 440                 445

Ile Gln Cys Phe Asp Trp Lys Arg Val Ser Glu Lys Lys Leu Asp Met
    450                 455                 460

Thr Glu Asn Asn Trp Ile Thr Leu Ser Arg Leu Ile Pro Leu Glu Ala
465                 470                 475                 480

Met Cys Lys Ala Arg Pro Leu Ala Ser Lys Val Glu Ser Tyr
                485                 490
```

<210> 5
<211> 843
<212> DNA
<213> Glycine max

<220>
<221> unsure
<222> (476)

<220>
<221> unsure
<222> (657)

<220>
<221> unsure
<222> (703)

<220>
<221> unsure
<222> (712)

<220>
<221> unsure
<222> (789)

<220>
<221> unsure
<222> (843)

<400> 5
ttcactctca agcttcaagc atgactcctt tttacttcct cctatttgcc ttcatccttt 60
tcctctccat aaacttcttg atccaaacaa gaaggttcaa aaaccttcct ccgggaccat 120
tttctttccc tataatcgga aacctccacc aactcaagca acccctccac cgcacgttcc 180
atgccttatc acaaaaatat ggccctattt tctccctctg gttcggctcc cgttttgtcg 240
tcgtcgtttc gtcgccgctc gcggtgcaag aatgcttcac caagaacgac atcgtcttgg 300
ccaaccgccc tcacttcctc accggcaagt atatagggta caacaacacc accgtcgccg 360
tttcccccta cggcgaccac tggcgcaacc tccgccgcat catggcgctc gaggttctct 420
ccacccaccg gataaactcc ttcttggaaa atcggagggg acgaagatca tgaggntcgt 480
gcaaaagctt gctcgggact cgcgcaatgg gttcaccaaa gtagaactta aatccaggtt 540
ttcggagatg acatttaaca ctataatgag gatggtgtca gggaagaggt actatggtga 600
agactgtgat gtgagtgatg tacaggaagc aagcaattta gagagatcat taaagantgg 660
tgacgttagg aggggctaat aacctgggga ctcttggttt gcntggtggt tntttgatgg 720
ttggaaagag ctaaagagga tagtagagaa cgatcgttta caggaccatt gtagcatcta 780
tggaacacnt gcatacatga taatatctct gccacacaac acaccgatat aacgttaatc 840
atn 843

<210> 6
<211> 141
<212> PRT
<213> Glycine max

<400> 6
Phe Leu Leu Phe Ala Phe Ile Leu Phe Leu Ser Ile Asn Phe Leu Ile
1 5 10 15

Gln Thr Arg Arg Phe Lys Asn Leu Pro Pro Gly Pro Phe Ser Phe Pro
20 25 30

Ile Ile Gly Asn Leu His Gln Leu Lys Gln Pro Leu His Arg Thr Phe
35 40 45

His Ala Leu Ser Gln Lys Tyr Gly Pro Ile Phe Ser Leu Trp Phe Gly
50 55 60

```
Ser Arg Phe Val Val Val Val Ser Ser Pro Leu Ala Val Gln Glu Cys
 65                  70                  75                  80

Phe Thr Lys Asn Asp Ile Val Leu Ala Asn Arg Pro His Phe Leu Thr
                 85                  90                  95

Gly Lys Tyr Ile Gly Tyr Asn Asn Thr Thr Val Ala Val Ser Pro Tyr
            100                 105                 110

Gly Asp His Trp Arg Asn Leu Arg Arg Ile Met Ala Leu Glu Val Leu
        115                 120                 125

Ser Thr His Arg Ile Asn Ser Phe Leu Glu Asn Arg Arg
    130                 135                 140
```